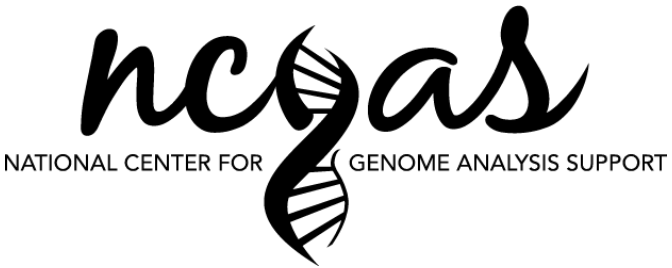
Compute resources available to the research community for microbiome analysis

Bhavya Papudeshi, Sheri Sanders, Carrie Ganote, Thomas G. Doak



NCGAS is an NSF-funded center tasked with assisting biologists in gaining access to computational resources needed to analyze genomic data (it's free!). This is done by providing:

- Access to compute resources to run their analyses
- Bioinformatics support- select pipelines/programs that best analyze the data to address research questions
- Training through workshops, sharing developed pipelines, online courses, and materials
- Gateways like Galaxy, GenePattern and genome browsers

ncgas
NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

Why should you care?

- Attending a bioinformatics workshop that has asked you to install a couple of programs beforehand

"This software will help you manage stress as long as you don't try to install it."

https://www.glasbergen.com/ngg_tag/manage-stress/

ncgas
NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

Jetstream

Jetstream, a cloud computing environment

ncgas
NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

Jetstream

Jetstream login | Request account

- Jetstream is an XSEDE (The Extreme Science and Engineering Discovery Environment) resource. XSEDE is an NSF project that makes available cyberinfrastructure to share data, run analysis and provides the expertise.

XSEDE

- Like cloud computing environments, you can spin up a virtual machine.

aws

Google Cloud

Jetstream, a cloud computing environment

- Virtual Machines are similar to a laptop, but computation is done elsewhere
- Users have root privileges, unlike high performance clusters
- Virtual machines help with the transition to command line use, software installation, and running analysis in the Linux environment
- Develop workflows or spin up pre-configured virtual machines, which can be shared and published between collaborators or publicly

Jetstream


Indiana University and Texas Advanced Computing Center (TACC) host the XSEDE cloud environment called Jetstream
jetstream

NCGAS use cases for Jetstream

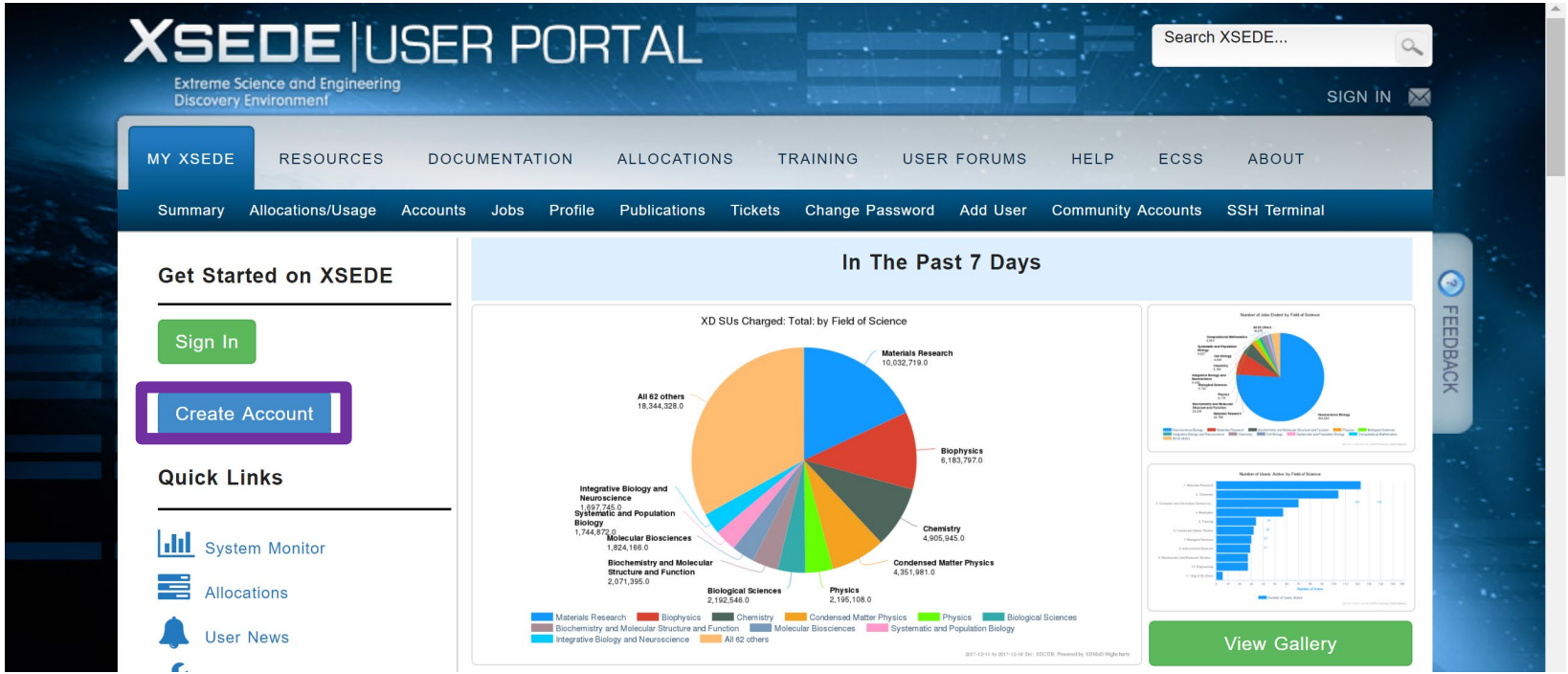
- Teaching workshops, was able to give all 50 workshop participants the exact same R setup, including all packages they needed.
- Used to provision a bioinformatics course at Bethane Cookman University.
- A prebuilt genome browser that eliminates much of the command line necessary to get started, to visualize data you just have to upload the data.
- MATLAB at no cost.
- Genomics toolkit image available with more than 30 packages already installed.
- Specialized bioinformatics pipelines- such QIIME2, PacBio, Galaxy service also available.

Getting access to Jetstream

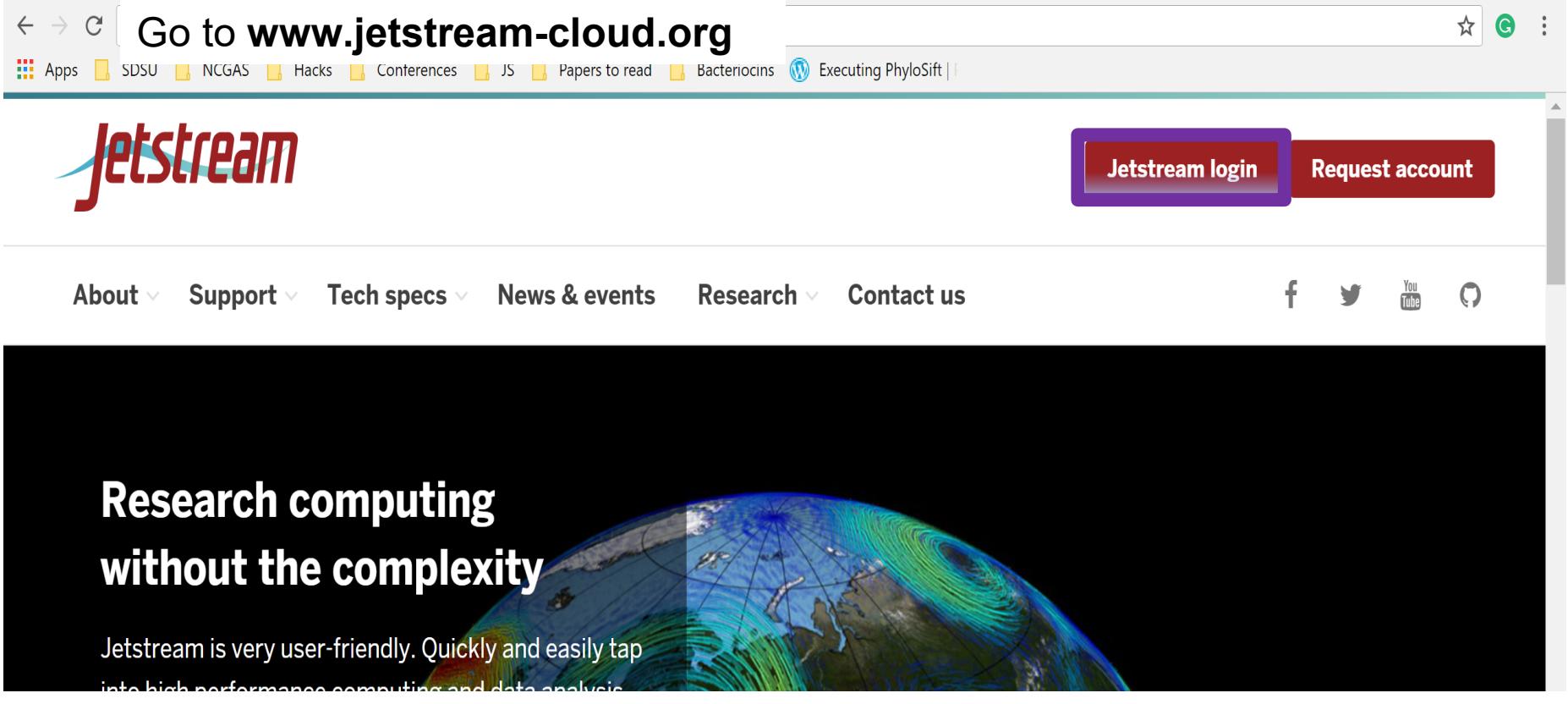
ncgas
NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

Step 1- Getting an XSEDE account

First –an XSEDE account, create one at https://portal.xsede.org



Jetstream, a cloud computing environment

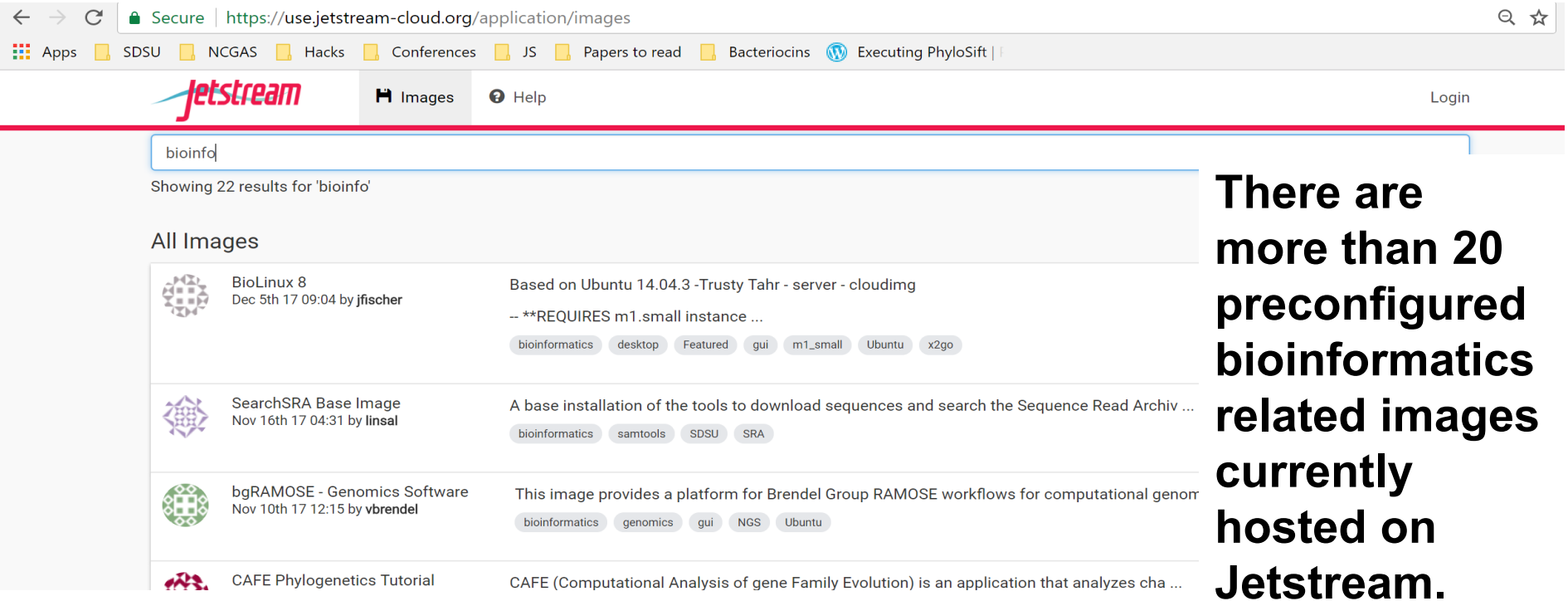


Step 2- Selecting a Jetstream allocation

First you will need a Jetstream allocation to get started,

1. Request for a Jetstream trial access- limited resources http://wiki.jetstream-cloud.org/Jetstream+Trial+Access+Allocation
2. Contact NCGAS (help@ncgas.org) to get on our allocation – trial access as well but more resources than Option 1
3. Request for your own Jetstream allocation- http://wiki.jetstream-cloud.org/Jetstream+Allocations

Preconfigured biology based VM's on Jetstream



There are more than 20 preconfigured bioinformatics related images currently hosted on Jetstream.

Configuring a Virtual Machine on Jetstream

Don’t see the software or workflow necessary for your research

Login ➡ startup a Ubuntu/CentOS image ➡ install software ➡ save/publish

- Install all the tools necessary for the workflow to the virtual machines
- Click on Image-> add image name and description, tags -> Next
- Type in the version number, base image used -> Next
- Access: public, specific users or private -> Submit

Generates a ticket to help@xsede.edu and wait for Jetstream team to image it for you.

Microbiome tools on preconfigured virtual machines

ncgas
NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

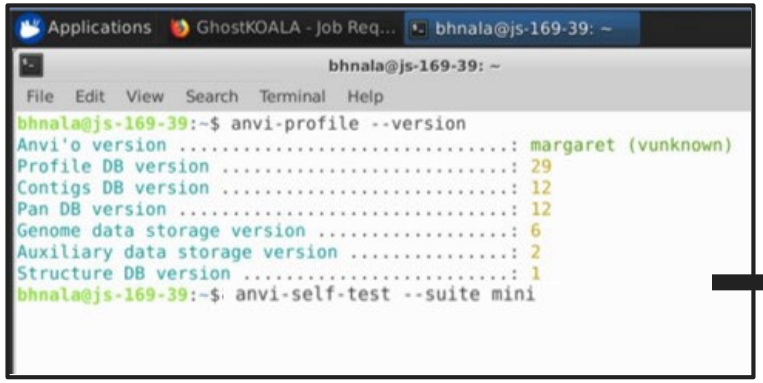
Microbiome related Jetstream images

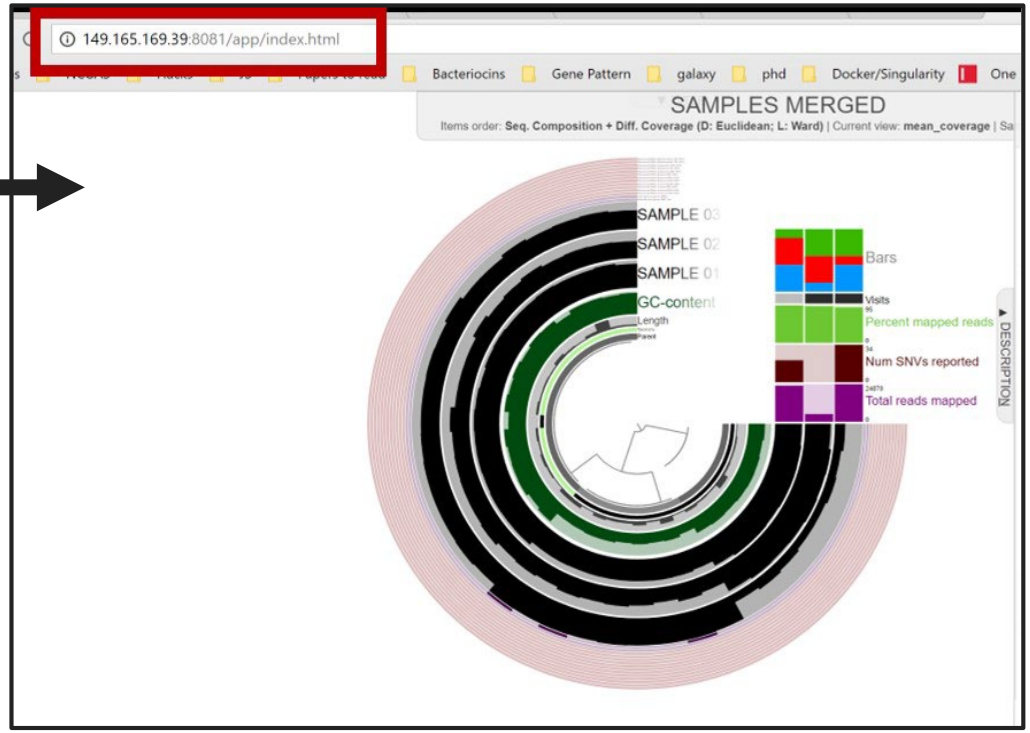
- **Anvi'o** - Anvi'o is an open-source, community-driven analysis and visualization platform for 'omics data. Link here
- **QIIME2 toolkit** - Quantitative Insights into Microbial Ecology is an open-source bioinformatics pipeline for performing microbiome analysis from raw DNA sequencing data. Link here
- **MEGAN** - "MEtaGenome ANalyzer" is a computer program that allows optimized analysis of large metagenomic datasets. Link here
- **Rstudio for NCGAS**- Set up and installed packages for various NCGAS workshops. Includes many genomic and mapping R packages and preloads three default users. Link here
- **PacBio SMRT Analysis** – PacBio analysis tools. Link here

There are more than 20 bioinformatics virtual machines

Start a preconfigured VM- Anvi'o

Start a preconfigured VM- Anvi'o





For more information -
[Running Anvio on Jetstream]

Acknowledgments

NCGAS Jetstream blog –

Getting Started on Jetstream
Setting up SSH keys
Setting up Globus on Jetstream
Setting up Volumes on Jetstream
Adding users to Jetstream image
Using Ansible to manage Jetstream images
Running Anvio on Jetstream
NCGAS maintained Jetstream images

Jetstream documentation - https://iujetstream.atlassian.net/wiki/spaces/JWT/overview

Contact NCGAS- help@ncgas.org

@ncgasiu
@ncgas

ncgas
NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

PITTSBURGH SUPERCOMPUTING CENTER

Jetstream

PERVASIVE TECHNOLOGY INSTITUTE


ncgas
NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT